DOCUMENT RESUME

ED 364 823 CG 025 151

AUTHOR Mount, Brian
TITLE Statistical Inference for the Admissions Officer.
PUB DATE Oct 93
NOTE 13p.; Paper presented at the Annual Meeting of the National Association of College Admission Counselors (49th, Pittsburgh, PA, October 7-10, 1993).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Descriptive (141)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Admissions Officers; Higher Education; *Marketing; Recruitment; *Statistical Inference

ABSTRACT

This paper, presented at a conference of college admissions counselors, attempts to provide a brief overview of descriptive and inferential statistics for college admissions officers, in the hopes that it will encourage these admissions personnel to question assumptions more critically. The paper begins by defining statistics, specifically inferential statistics (the science of interpreting data to make predictions, estimates, or decisions). It goes on to explain why statistics are useful to college admissions officers. Variables are explained and the distinction between correlational and experimental research is discussed. The distinction between independent and dependent variables also is described. Measurement scales are discussed as they relate to different types of variables (nominal, ordinal, interval, and ratio). Relations between variables are explored and the importance of those relations is considered. Samples and populations are discussed. Basic features of relations between variables are examined, statistical significance is explained, and how levels of significance are calculated is discussed. Finally, a number of basic statistical terms are defined, including frequency tables, crosstabs, chi-square, correlations, testing for differences in means between independent samples, t-tests, multiple regression, and nonparametric statistics. The paper concludes with a brief section on computer-based statistics programs. (NB)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 364 823

STATISTICAL INFERENCE FOR THE ADMISSIONS OFFICER

Presented at the 49th Annual Conference of the National Association
of College Admissions Counselors
Pittsburgh, PA October 7-10, 1993

Brian Mount
Director of Recruitment and Admissions

BEST COPY AVAILABLE

CG025151

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

B. Mount

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

STATISTICAL INFERENCE FOR THE ADMISSIONS OFFICER

Presented at the 49th Annual
Conference of the National Association
of College Admissions Counselors
Pittsburgh, PA October 7-10, 1993

Brian Mount
Director of Recruitment and Admissions

ABSTRACT

The word "statistics" frightens many individuals for no reason other than an unfamiliarity with the basics of statistical inference. Too many admissions officers do not critically evaluate their marketing research efforts. Thousands of dollars are spent on hunches or on simple tabulations of gross amounts of data without careful analysis of underlying relationships. Consultants are often hired to produce such analyses -- often at considerable expense and without producing any useful analyses because the right questions were not asked. This paper attempts to provide a brief overview of descriptive and inferential statistics and will encourage session participants to question assumptions more critically.

WHAT IS STATISTICS?

When most people hear the word *statistics*, they usually think of an assemblage of numbers that summarize and describe data. This is indeed one definition of the field of statistics known as **descriptive statistics** -- the science of summarizing or describing data. Descriptive statistics include such common tests for significance as t-Tests for independent and correlated samples (i.e., evaluating the differences between mean values for two groups). Descriptive statistics are useful in summarizing large amounts of data.

Of more use, particularly to admissions officers, is the study of **inferential statistics** -- the science of interpreting data to make *predictions*, *estimates*, or *decisions*. One of the most common statistical measurements utilized in this category is that of regression analysis, or more typically, multiple regression analysis, where the intent is to determine the relationship between an independent (or predictor) variable and a dependent (or criterion) variable. We can view inferential statistical analysis as the key enabling us to unlock the hidden relationships in data.

The word statistics is utilized throughout this paper without specifying which of the two definitions above are being drawn upon. In most instances, and unless noted otherwise, when we utilize the term statistics, we are referring to the process of statistical inference.

WHY ARE STATISTICS USEFUL?

Perhaps the most basic reason why we should familiarize ourselves with basic statistical inference is the fact, mentioned above, that statistics can help in briefly summarizing large amounts of data. This is particularly important if you have a large amount of data to summarize in only a few paragraphs for the President or CEO.

Second, statistics are useful in confirming or denying "gut" feelings. One of the papers attached to this brief outline is appropriately entitled "Enrollment Projections: Combining Statistics and Gut Feelings." This paper outlines an attempt to replace admissions "hunches" with more reliable models to predict enrollment for budgetary and academic planning purposes.

Third, statistics can be useful in justifying budget requests and allocations for admissions officers. If a relationship can be shown to exist between expenditures in a particular area, e.g., advertising, and applications, then statistical analysis can be invaluable in justifying budgetary requests.

Fourth, and finally, one of the best reasons to familiarize ourselves with statistics is the fact that other people use, and misuse them, constantly. Take for example, the two advertising claims cited as examples of statistical misuse in A.K. Dewdney's *200% of Nothing*:

> • "You'd have to eat four bowls of raisin brain to get the vitamin nutrition in one bowl of Brand X." (The phrase "the vitamin nutrition" doesn't necessarily mean "all the vitamins." It could mean that Brand X has

more of one minor vitamin and less of all the important vitamins."

• "Our antacid consumes forty-seven times its weight in excess stomach acid." ("Forty-seven times" may or may not be impressive. By itself, this claim is no more meaningful than the statement, "Bill weighs forty-seven times as much as his dog." If you don't know Bill's age, weight, or the weight of Bill's dog, you have no reference point from which to judge whether Bill is overweight or not. Likewise, if you don't know the weight of the antacid tablet -- or of the weight of the "excess" stomach acid -- then you can make no judgment on the validity of this impressive claim.

You have to be able to not only know what statistics to use and when, but you need to know enough to question when they are misused to support an issue or a claim.

For the beginner, recommended introductory textbooks include: Blalock (1979), Dietrich and Kearns (1989), and Kachigan (1988). More advanced statistical users may wish to refer to Hays (1988) and Kendall and Stuart (1979).

WHAT ARE VARIABLES?

Variables are things that we measure, control, or manipulate in research. In examining admissions statistics, students may become a variable. We may not be able to control or manipulate them (much to our faculty's chagrin), but we can certainly measure certain aspects about them. Variables may differ in many respects, most notably in the role they are given in our research and in the types of measures that can be applied to them.

CORRELATIONAL VS. EXPERIMENTAL RESEARCH

Most empirical research belongs to one of these two general categories. In *correlational* research we do not attempt to influence any variables, but only seek to measure them and look for relations (or correlations) between some set of variables, such as financial aid awards and the likelihood of enrollment.

In *experimental* research, we seek to manipulate some variables and then measure the effects of this manipulation on other variables. For example, using the same variables we listed above -- financial aid awards and the likelihood of enrollment -- we could design an experimental design whereby we would purposely vary the amount of financial aid awards to selected students in an attempt to measure, experimentally, the impact of such awards upon their enrollment. Data analysis in experimental research also revolves around calculating "correlations" between variables. Experimental data, however, may potentially provide qualitatively better information for only experimental data allows us to conclusively *demonstrate causal* relations between variables. If we vary the amount of financial aid awards and find that the likelihood of enrollment is likewise altered, then we may conclude that financial aid influences enrollment. It is most important here, however, to note that *we can only infer causation based upon the theory that we advance* (the theory being that financial aid awards influence enrollment behavior). Correlational data does not conclusively prove causality. The step of inferring causality from the data we obtain is the inferential component of statistics and one of the mostly widely misused.

DEPENDENT & INDEPENDENT VARIABLES

Independent variables are those that we are able to manipulate, while *dependent* variables are those that we simply measure or register. This distinction may appear confusing to many as "all variables depend upon something." Once you acquaint yourself to this distinction, however, the importance between the two sets of variables becomes empowering.

To help clarify the distinctions noted above, consider that the terms *dependent* and *independent* variable are found mostly in experimental research where one or more variables are manipulated and in this sense are "independent" from the initial response. Other variables are expected to be "dependent" upon the

manipulation or experimental conditions. In other words, they are "dependent" upon the response.

In admissions, we don't have the opportunity to often conduct experimental research. It is more likely that we will assign students to "experimental groups" based upon some preexisting properties they possess. For example, if in an analysis, young males are compared with young females regarding their propensity to enroll once extended an offer of admission, *gender* or *sex* could be called the *independent* variable and the propensity to enroll the *dependent* variable -- the propensity to enroll is "dependent" upon a person's gender.

MEASUREMENT SCALES

Variables differ in how well they can be measured, i.e., in how much information their measurement scale can provide. There is obviously some error involved in very measurement, which determines the amount of information that we can obtain. The variable "gender" utilized in the example above is one that we can definitively measure. Nevertheless, it may be coded wrong or the information supplied to us by the student may be inaccurate. Likewise, the propensity to enroll can be easily measured by whether or not a student enrolls in a given time frame. But what about those students who elect to defer enrollment? They don't enter in the time frame we are measuring, yet they may become students at a later time. Another factor that determines the amount of information that can be provided by a variable is its "type of measurement scale." Variables may be classified according to one of four types: (a) *nominal*; (b) *ordinal*; (c) *interval*; or (d) *ratio*.

(a) *Nominal* variables allow only for qualitative classification. In other words, they can be measured only in terms of whether the individual items belong to some distinctly different categories, but we cannot quantify or even rank order those categories. Typical examples of *nominal* variables used in admissions are gender, race, city, etc. We can for example, say that two individuals are different in terms of the variable race, but we cannot say which one "has more" of the quality represented by the variable race.

(b) *Ordinal* variables allow us to rank order the items we measure in terms of which has less and which has more of the quality represented by the variables. They do not, however, allow us to say "how much" more. A typical example of an ordinal variable utilized in admissions is socioeconomic status of families. We know, for example, that an *upper-middle* income families are regarded as being in a higher category than *middle income families*, but we cannot say that they are, for example, twice as high or 100% higher. Clearly, *ordinal* variables provide us more information than *nominal* variables but they still do not allow us to say "how much" more information we have. For this purpose, we need *interval* variables.

(c) *Interval* variables allow us the ability to not only rank order the items that we are seeking to measure, but they also allow us the opportunity to quantify and compare the differences between the variables. Income, as measured in dollars, constitutes an interval scale. We can not only say that income of $50,000 annually is higher than an income of $25,000 annually but we can say that in increase from $30,000 to $60,000 annually is twice as much as an increase from $30,000 to $45,000. The width of the intervals used is also of prime importance when evaluating interval variables and *depends upon the variable being measured*. A $500 difference between housing prices in the $200,000+ range may be insignificant but a $500 difference in tuition between two competing schools may be significant.

(d) *Ratio* variables are very similar to interval variables but they have one important distinction, they feature an absolute zero point which allows for statements to the effect that "x is 2 times more than y." Most census data is obtained by counting the various kinds of people in each category and dividing by the population base: % urban, % unemployed, etc. While interval scales do not have the ratio property, most statistical data analysis procedures do not distinguish between the two measurement scales and distinctions between them, while interesting (Coleman 1964), are of little consequence to us here.

RELATIONS BETWEEN VARIABLES

Regardless of the type of variable, two or more variables are related if in a sample of observations the values of those variables are distributed in a consistent manner. In other words, variables are related if their values systematically correspond to each other for these observations. For example, *gender* and *propensity to enroll* would be considered to be related if in most females there was a high *propensity to enroll* and in most males a low *propensity to enroll*, or vice versa. *College GPA* is related to *SAT test scores* if students with higher SAT test scores perform better in college. Note that we're not making a judgment about the quality of the relationship here or whether other variables may have an impact as well, we're merely talking about the relationship between variables.

WHY RELATIONS BETWEEN VARIABLES ARE IMPORTANT

The goal of any analysis is to find relations between variables. The philosophy of science as expressed by Hume and others teaches us that there is no other way of representing "meaning" except in terms of relations between some quantities or qualities: either way involving relations between variables. Consequently, the advancement of science must always involve finding new relations between variables. Correlational research involves measuring such relations in the most straightforward manner. Experimental research is similar in this respect. It is the role of statistics to help us evaluate the relations between variables. *All of the hundreds of statistical procedures available to us are nothing more than means of evaluating various kinds of inter-variable relations.*

SAMPLES AND POPULATIONS

A *population* is the collection of <u>all</u> objects of particular interest to us in our measurement. A *sample* is a <u>set</u> of objects or data taken from all possible objects in the population. A sample is but a representation or subset of the population and is consequently smaller than the population. Statistical inference deals with *samples* which allow us to study but a portion of the overall population and generate estimations, predictions, or conclusions regarding relationships which should describe the population.

Sampling procedures are based upon the premise that the samples drawn from the population are random and free from bias. Random, independent samples help ensure that were the experiment to be replicated, the same relationships would likely be found to exist.

BASIC FEATURES OF RELATIONS BETWEEN VARIABLES

There are two properties common to every relationship between variables: (a) the magnitude or size of the relationship; and (b) the reliability or truthfulness of the relationship.

(a) *Magnitude (or "size").* If in the example we used previously, every young male student was found to have a higher *propensity to enroll* than any female in the study, we could say that the magnitude of the relation between the two variables (*gender* and *propensity to enroll*) is very high. In other words, knowing the value of one variable, we could predict a value of the other, e.g., a female student could be predicted to have a lower propensity to enroll.

(b) *Reliability (or "truthfulness").* The *reliability* of a relation is less intuitive than the concept of magnitude. It pertains to the "representativeness" of the result found in a specific study (or sample) to the entire population or how certain we are that an inference is correct. In other words, it says how probable it is that a similar relation would be found if the experiment was repeated with other samples (cadres of students) drawn from the same population (e.g., all students at the university).

As a rule, we are almost never interested in what is going on in the sample we've drawn from (an example would be an incoming freshman class), but we're interested in the sample to the extent that it can help us to understand something about the entire population (all incoming students in future years). The reliability of a relation between variables observed in our sample can be

quantitatively estimated and represented using a standard measure known as the *significance level or p level.*

STATISTICAL SIGNIFICANCE

The statistical significance of a result is an estimated measure of the degree to which it is "true" or "representative of the population." The value of the *p* level represents a decreasing index of the reliability of a result. The higher the *p* level, the less we can believe that the observed relation between variables in the sample is a reliable indicator of the relation between the respective variables in the population.

Specifically, the *p* level represents the probability of error that is involved in accepting our observed result as valid. For example, the *p* level of .05 commonly used in much social research in admissions indicates that there is a 5% probability that the relationship between the variables found in our sample is not valid or could have happened by chance. Assuming that in the population of all students, we repeated our relating financial aid to the propensity to enroll, we could expect that approximately one in every twenty replications there would be a result in which the relation between the variables would be equal or stronger than in ours. A *p* level of .05 is generally regarded as a lower-bound acceptable limit. Some social research utilizes a .1 percent *p* level -- meaning that the researchers are willing to accept a 10% chance that their results are a "fluke." Always examine the level of significance utilized in a study to provide a quick benchmark regarding the reliability of the relationship being studied.

HOW DO YOU DETERMINE THAT A RESULT IS TRULY SIGNIFICANT?

With the above caution in mind, however, it is important to note that there is really no way to avoid some arbitrary decision-making as to what level of significance should be treated as really significant. In practice, the final decision usually depends on whether the outcome was predicted prior to the act (*a priori*) or only found after the experiment or study was concluded (*post hoc*).

Typically, results that yield $p \leq .05$ are considered borderline statistically significant. Results that are significant at the $p \leq .01$ level are commonly considered statistically significant and $p \leq .005$ or $p \leq .001$ levels are often referred to as highly significant. Overall, however, you should keep in mind that these classifications represent highly arbitrary numbers. In testing the efficacy of a drug, as an example, pharmaceutical manufacturers must set extremely high levels of significance in examining whether any side-effects associated with the drug may be due to chance or a result of the drug. When dealing with enrollment predictions, you can obviously accept a much higher level of significance. The real question to be considered, however, is how much room for error will the board of trustees allow you?

STATISTICAL SIGNIFICANCE AND THE NUMBER OF ANALYSES PERFORMED

The more analyses you perform on a particular data set, the more results will "by chance" meet the conventional significance level. Some statistical methods that involve many comparisons, and thus a good chance for errors, include some correction or adjustment for the total number of comparisons. Such methods often involve the calculation of "degrees of freedom". The smaller the number of degrees of freedom associated with a particular statistic, the more variable its sampling distribution will be. Likewise, higher degrees of freedom (*d.f.*) imply a greater likelihood that the sample will approximate the population. Ultimately, it is up to you to evaluate the reliability of your findings, particularly unexpected findings.

STRENGTH VS. RELIABILITY OF RELATIONSHIPS BETWEEN VARIABLES

Generally, in a sample of a given size, the larger the magnitude of the relationship between variables, the more reliable the relationship (higher significance level).

The magnitude and significance of a relation appear to be closely related, and it is possible to calculate the significance from the magnitude and

vice-versa. This is only true, however, if the sample size is kept constant because a relationship could be either highly significant or not significant at all depending upon sample size. If there are very few observations, then there are also respectively few possible combinations of the values of the variables and thus the probability of obtaining by chance a combination of those values (remember the significance level we discussed earlier) is relatively high.

Consider the following example. If we are interested in two variables (*gender*: male/female and *propensity to enroll*: high/low) and there are only four subjects in our sample (two males and two females), then the probability that we will find, purely by chance, a 100% relation between the two variables can be as high as one-eighth. Specifically, there is a one-in-eight chance that both males will have a high *propensity to enroll* and both females a low *propensity to enroll*, or vice versa. If, on the other hand, our sample consisted of one hundred students, or perhaps the entire incoming class, the probability of obtaining such an outcome by chance would be so small as to be practically zero. The smaller the sample size, the more likely it is that we will obtain erroneous results. The larger the sample, the greater the likelihood that the results we receive are indicative of the entire population. Technically speaking, the probability of a random deviation of a particular size (from the population mean), decreases with the increase in the sample size.

SMALL RELATIONS CAN BE PROVEN SIGNIFICANT ONLY IN LARGE SAMPLES

Even if the sample we draw is "perfectly representative" of the overall population, the effect will not be statistically significant if the sample is small. Conversely, if a relationship is "objectively" very large (i.e., within the population), then it can be found to be highly significant even if only a very small sample is taken.

CAN "NO RELATIONSHIP" BE A SIGNIFICANT RESULT?

The smaller the relationship between variables, the larger the sample size that is necessary to prove it significant. The necessary minimum sample size increases as the magnitude of the effect to be demonstrated decreases. The magnitude of the effect is approaching zero, the necessary sample size to conclusively prove a relationship approaches infinity. In other words, if there is almost no relationship between two variables, then the sample size must be almost equal to the population size which is assumed to be infinitely large. *Statistical significance represents the probability that a similar outcome would be obtained if we tested the entire population.*

MEASURING THE MAGNITUDE (STRENGTH) OF RELATIONSHIPS BETWEEN VARIABLES

Almost all measures of measures of the magnitude of relationships between variables adhere to one general principles: They attempt to somehow evaluate the observed relation by comparing it to the "maximum imaginable relation" between those specific variables. In other words, measurements of the strength of relationships compare "what is common between variables" to "what potentially could have been common if the variables were perfectly related."

As the ultimate goal of most statistical tests is to evaluate the relationships between variables, most statistical tests follow the general format that was explained in the previous paragraph. More formally, statistical tests represent a ratio of some measure of the differentiation common in the variables in question to the overall differentiation of those variables. This ratio is usually called a ratio of *explained* variation to *total* variation. The term *explained variation* means that part of the variation in one variable is explained by the variation in another. The underlying conceptualization of what actually is happening -- why the variation in one variable is "explained" by another -- is the province of the researcher and is the art of interpretation in statistical analysis.

CALCULATING LEVELS OF SIGNIFICANCE (alternative ($\mathcal{H}a$) and null ($\mathcal{H}_o$) hypotheses)

Once a measure of a relationship is calculated, the next question you would ask is how significant is this relationship? Like so much in life, the answer is "it depends."

Specifically, the significance depends primarily upon the sample size. In very large samples, even very small relations between variables will be significant whereas in very small samples even very large relations cannot always be considered significant (reliable). We are interested in knowing how likely it is to obtain a relationship of a given magnitude or larger, from a sample of a given size, assuming that there is no such relationship between those variables in the population. This function would provide us with a significance (p) level and it would tell us the probability of error involved in rejecting the idea that the relationship in question does not exist in the population.

In a test of hypothesis, the hypotheses of interest are given special names. In the example cited above (that there is no relationship in the population), the hypothesis is known as the *null hypothesis* ($\mathcal{H}_o$). The purpose of conducting a statistical evaluation is to determine whether the null hypothesis is unreasonable. The alternative hypothesis provides an opposing conjecture to that given in the null hypothesis. Statistical tests are conducted to determine whether the *alternative hypothesis* ($\mathcal{H}_a$) is supported. The alternative hypothesis is often referred to as the research hypothesis. The statistic we calculate to determine whether the null hypothesis is rejected is known as a test statistic (z, t, etc.). It is important to note, that tests are conducted to reject the null hypothesis and accept the alternative hypothesis. Rejecting the null hypothesis when $\mathcal{H}_o$ is indeed true is known as a Type I error. Alternatively accepting the null hypothesis when $\mathcal{H}_o$ is false is known as a Type II error.

NORMAL DISTRIBUTION

You've probably heard the term "normal distribution" before. The normal distribution is important because in most cases, it best approximates the general nature of reality. The exact shape of the normal distribution (typically shown as the characteristic "bell curve") is defined by only two parameters: the *mean* and the *standard deviation*.

The mean is a measure of central tendency and mathematically is computed by dividing the sum of all the measurements by the total number of measurements in the data set. Along with the mean, another less commonly utilized measure of central tendency is the median -- the number which divides a set of data exactly in half (smallest to largest). The mean may overstate the central tendency of a sample if there are large outliers in the sample. In such cases, the median may be a more appropriate measure of central tendency. Take housing prices for example. The average price of housing in an area is always reported as a median figure, not a mean figure. The median figure indicates that half the homes sold below this figure, half above. If there were a small number of very expensive homes, calculation of a mean price might distort the "average" price paid for housing (make it considerably higher).

A characteristic property of the normal distribution is that 68% of all its observations fall within a range of $\pm$ 1 *standard* deviation from the mean. The standard deviation is the square root of the arithmetic mean of the squared deviations from the mean. A range of $\pm$ 2 standard deviations includes 95% of the data in a normal distribution. In other words, in a normal distribution, observations that have a standardized value of -2 or +2 have a relative frequency of 5% or less. (Standardized value means that a value is expressed in terms of its difference from the mean, divided by the standard deviation.)

USING THE NORMAL DISTRIBUTION IN STATISTICAL REASONING

If a sample size is large enough, the results of repeated experiments are "normally distributed," and thus knowing the shape of the normal curve, we can precisely calculate the probability of obtaining "by chance" outcomes representing various levels of deviation from the hypothetical population mean of 0. If such a calculated

probability is low that it meets the previously accepted criterion of statistical significance, then we have but one choice: conclude that our result gives a better approximation of what is going on in the population than does the "null hypothesis" – in other words, we reject $\mathcal{H}_o$.

Not all but most of the test statistics you will have the opportunity to employ are either based on the normal distribution directly or on distributions that are related to, and can be derived from normal such as t, F, or χ^2. Typically, these tests require that the variables analyzed are themselves are normally distributed in the population.

When variables are not normally distributed, we have but two options. First, we can utilize Nonparametric tests but these tests are less flexible and less powerful in terms of the conclusions they provide. Alternatively, we can still utilize the normal distribution-based test if we only make sure that the size of the samples we draw are large enough for as sample size increases, the shape of the sampling distribution approaches that of the normal distribution.

BASIC STATISTICS

The following paragraphs touch briefly upon some of the basic statistics which you might come across in your efforts to address questions of statistical inference.

FREQUENCY TABLES

Basic statistics include the grouping of data into frequency tables. A frequency table will show the number of observations that fall into different categories or ranges across the values of the respective variable.

CROSSTABS

Crosstabulation of data is a combination of *two* frequency tables arranged such that each cell in the resulting table represents a unique combination of specific values of crosstabulated variables. Crosstabulation allows us to examine frequencies of observations that belong to specific categories on more than one variable. The simplest form of the crosstabulation procedure is the 2 x 2 table where each variable has only two distinct choices. Take the following example where students are grouped by the variable *gender* in regards to their preference for either School A or School B:

	School A	School B
Male	10 (33%)	20 (67%)
Female	20 (67%)	10 (33%)
	30 (100%)	30 (100%)

Each cell represents a unique combination of values of the two crosstabulated variables and the numbers in each cell tell us how many observations fall into each combination of values. In general, the table tell us that more males than females prefer school B and vice-versa.

CHI-SQUARE (χ^2)

The Pearson Chi-square (χ^2) is the most common test of significance between categorical variables. The Chi-square statistic is based upon the fact that we can determine the expected frequencies and then compare the *expected* outcomes with those we actually obtain. You will need to utilize a Chi-square table and adjust for the appropriate degrees of freedom to determine whether relationships are significant. As a general rule, the larger the Chi-square value you obtain, the more significant the relationship.

The only assumption underlying the use of the χ^2 statistic is that the expected frequencies are not very small. When expected frequencies in the cells fall below 5, for example, those probabilities cannot be further estimated with sufficient precision.

CORRELATIONS

Correlation is a measure of the relationship between two or more variables. The value of a correlation does not depend upon the specific measurements used. *Proportional* means *linearly related*, that is the correlation is high if it can be "described" by a straight line. You've probably seen straight lines through scatterplots of data elements. This line is called the *regression line* or *least squares line*, because it is determined by calculating that the sum of the squared distances of all the data points from the line is the smallest possible number.

The slope of the regression line is indicative of the magnitude (strength) of the correlation. Simple linear correlation is measured by the strength of a statistic known as the Pearson correlation or Pearson r. The square of the correlation coefficient, known as the *coefficient of determination* (r^2), will represent the proportion of common variation in the two variables.

t-TESTS

The *t*-test is the most commonly used method to evaluate differences in means between two groups. The *t*-test can be used when the sample sizes are very small. *t*-tests can be used in evaluating the differences between independent samples as well as dependent or *correlated* samples.

TESTING FOR DIFFERENCES IN MEANS BETWEEN INDEPENDENT SAMPLES

When testing for the differences between means in a sample, let's say, for example, that we have ten years of data which show the average ACT score for entering freshmen to be 28 with a standard deviation of 3. For the last incoming class of 75 students, the ACT score was 29 with a standard deviation of 4. Can you tell your President with any confidence that the ACT scores at your institution in this last year were significantly different from prior years, or was this due strictly to chance?

In this particular and very simplified instance of comparing the means between two samples, we have to first assess whether the size of the sample is large enough to utilize a normal statistic (z) or an approximation of the normal statistic (t-test). You should always be wary of rules of thumb, but here's one anyway. For sample sizes of 100 or greater, you can be confident that the normality assumption can be relaxed and you can utilize the z-test statistic. For samples of between 50 to 100, you may also utilize the z-test if you are confident that the sample is from a relatively normally distributed population. When the sample size is ≤ 30, however, always utilize the t-test -- there are not enough cases in the sample to indicate the form of the population distribution.

MULTIPLE REGRESSION

The general purpose of multiple linear regression is to determine the relationship between several independent or predictor variables and a dependent or criterion variable. For example, we might wish to know what factors influence a student's *propensity to enroll* (the dependent variable) besides *gender*. We might also wish to include *socioeconomic status, HS rank, whether a personal visit was made to campus*, etc.

The regression line is a straight line fit of the equation $Y = a + b * X$. Y, the dependent variable, can be expressed in terms of a constant (*a*)-- also known as the intercept -- and a slope (*b*) times the X variable. The slope (*b*) is also known as the regression or Beta ($\mathcal{B}$) weight.

In the above equation, the $\mathcal{B}$ weights represent the independent contributions of each independent variable to the prediction of the dependent variable. This type of correlation is also known as *partial correlation*. The regression line expresses the best prediction of the dependent variable (Y), given the independent variables (X). There are usually substantial variation of the observed plotted points around a fitted regression line. The deviation of a particular point from the regression line is called the *residual value*.

The smaller the variability of the residual values around the regression line relative to the overall variability, the better is our prediction. For example, if there is no relationship between the X and Y variables, then the ratio of the residual

variability of the Y variable to the original variance is equal to 1.0. If X and Y are perfectly related then there is no residual variance and the ratio of variances would be 0. In most cases, the ratio would fall somewhere between these extremes, that is between 0 and 1. 1 minus this ratio is referred to as R^2 or the *coefficient of determination*. If we have an R^2 of 0.4, then we know that the variability of the Y values around the regression line is 1-0.4 times the original variance. In other words, we have explained 40% of the original variability, and are left with 60% residual variability. Remember, regression analysis *does not infer causality*, only that a relationship may be shown to exist. It is up to the researcher to imply the existence of causality from the data at hand.

The degree to which two or more predictors (independent or X variables) are related to the dependent (Y) variable is expressed in the correlation coefficient R^2, which is the square root of R. In multiple regression, R can assume values between 0 and 1. To interpret the direction of the relationship between variables, one looks at the signs (plus or minus) of the regression or . $\mathcal{B}$ weights. If a $\mathcal{B}$ weight is positive, then the relationship of this variable with the dependent variable is positive. If the $\mathcal{B}$ weight is negative, then the relationship is negative. If the $\mathcal{B}$ weight is 0, then no relationship between the variables is shown to exist.

There are limitations to multiple regression analysis as well as different types of regression analysis (e.g., Stepwise, Logit). For a more complete description of the uses of multiple regression analysis than is permitted here the reader is referred to Blalock (1979) cited earlier.

NONPARAMETRIC STATISTICS

Some information in the general population is not normally distributed, consequently, it becomes necessary to consider the utilization of nonparametric statistics to account for such non-normal distributions. Examples of what are probably non-normal distributions include income, incidence rates of rare diseases, the # of car accidents. In such instances, we have to consider nonparametric statistics as alternatives to the following:

Alternatives to the t-test for independent samples -- Wald-Wolfowitz Runs test, Mann-Whitney U test, and Kolmogorov Smirnov two-sample test.

Alternatives to the t-test for dependent (correlated) samples -- Sign test and Wilcoxon matched pairs test.

Alternatives to correlation -- Spearman R, Kendall tau, Gamma, and Kendall coefficient of concordance.

COMPUTER BASED STATISTICS PROGRAMS

Happily, there are sophisticated PC-based software programs available today enabling anyone to become quickly acquainted with the power of statistical analysis.

A number of such programs exist. For the DOS or Windows environment, popular programs include STATISTICA™, Stanford Graphics♦, Systat for Windows♦, and SPSS for Windows♦. For the Macintosh environment, there are Systat™, SPSS♦, StatView™, SuperANOVA™, JMP™, Data Desk♦, and STATISTICA/Mac™.

Each of the above mentioned programs offers differing advantages regarding speed, price, flexibility, etc. Having tried a majority of the programs mentioned above, the author heartily recommends STATISTICA™ and STATISTICA/Mac™ as the clear favorite in terms of speed, flexibility, general ease of use, graphics capabilities and power. The STATISTICA™ programs (either DOS or Windows based) are <u>not</u> among the least expensive of the options available, but offer an excellent alternative in many other respects.

Information on STASTISTICA is available from:

Statsoft, Inc.
2325 East 13th Street
Tulsa, OK 74104
(918) 538-4149

Information on StatView® & SuperANOVA is available from:

Abacus Concepts
1984 Bonita Avenue
Berkeley, CA 94704-1038
(510) 540-1949

Information on SAS® is available from:

SAS Institute Inc.
Software Sales Department
SAS Campus Drive
Cary, NC 27513
(919) 677-8000

Information on SPSS and SPSS for Windows is available from:

SPSS, Inc.
444 North Michigan Avenue
Chicago, IL 60611
(312) 329-3300
(800) 543-2185

REFERENCES

Blalock, H. M. (1979), *Social Statistics* (2nd ed.), New York, NY: McGraw-Hill, Inc.

Coleman, J. S. (1964), *Introduction to Mathematical Sociology*, New York, NY: The Free Press.

Dewdney, A.D. (1993), *200% of Nothing*, New York, NY: John Wiley & Sons.

Dietrich, F. H. II, and T. J. Kearns (1989), *Basic Statistics: An Inferential Approach*, San Francisco, CA: Dellen Publishing Company.

Hays, W. L., (1988), *Statistics* (4th ed.), New York, NY: CBS College Publishing.

Kachigan, S.K. (1986), *Statistical analysis: An interdisciplinary introduction to uivariate and multivariate methods*, New York, NY: Redius Press.

Kendall, M., and Stuart, A. (1979), *The advanced theory of statistics* (Vol. 2), New York, NY: Hafner Publishing Company.